PFMFind: a system for discovery of peptide homology and function

Aleksandar Stojmirović [1,2] *, Peter Andreae [2], Mike Boland [3],
Thomas William Jordan [4], Vladimir G. Pestov [1]

(1) Department of Mathematics and Statistics,
University of Ottawa,
585 King Edward Ave., Ottawa, ON K1N 6N5, Canada

(2) School of Mathematics, Statistics and Computer Science,
Victoria University of Wellington,
PO Box 600, Wellington, New Zealand

(3) Fonterra Research Centre,
Private Bag 11029, Palmerston North, New Zealand

(4) School of Biological Sciences,
Victoria University of Wellington,
PO Box 600, Wellington, New Zealand

Abstract

Summary:

Protein Fragment Motif Finder (PFMFind) is a system that enables efficient discovery of relationships between short fragments of protein sequences using similarity search. It supports queries based on score matrices and PSSMs obtained through an iterative procedure similar to PSI-BLAST. PSSM construction is customisable through plugins written in Python. PFMFind consists of a GUI client, an algorithm using an index for fast similarity search and a relational database for storing search results and sequence annotations. It is written mostly in Python. All components communicate between themselves using TCP/IP sockets and can be located on different physical machines. PFMFind is available for UNIX and Windows platforms.

Availability:

PFMFind is freely available (under a GPL licence) for download from the web site of the Centre for Biodiscovery, Victoria University of Wellington, http://www.vuw.ac.nz/biodiscovery/publications/centre/pfmfind.aspx

Contact:

astojmir@uottawa.ca

*to whom correspondence should be addressed

Introduction

The biological functions of proteins are as much a function of particular motifs of peptide sequence as they are of the overall protein structure. It is of interest to the biologist to search for examples of convergent motifs as they are likely to indicate a functional role. While many approaches exist for finding longer sequence motifs (50 amino acids or more), finding relationships between short fragments (3–18 amino acids long) of full protein sequences also promises great rewards in understanding novel aspects of protein structure and function. These relationships might be evolutionary in origin or might arise by convergence, that is, by acquisition of the same biological function in evolutionarily distant species.

Finding short motifs presents significant challenges because many of the apparent relationships between short fragments could have arisen by chance and thus have no functional significance. Furthermore, most widely available tools for sequence database search and motif finding were designed with longer motifs in mind. For example, Watt and Doyle (11) recently observed that the NCBI BLAST (1) family of programs, the best known set of tools for searching biological sequence datasets, is not suitable for identifying shorter sequences with particular constraints and proposed a pattern search tool to find DNA or protein fragments that match a given sequence or a pattern exactly. This paper outlines the Protein Fragment Motif Finder (PMFind), a new tool that uses database search to identify the conserved short peptide motifs of a query sequence and associates them with the available functional annotations.

Overview

The PFMFind system consists of three major components: a search engine for fast similarity search of datasets of short peptide fragments called FSIndex, a relational database, and the PFMFind GUI (graphical user interface) client (Figure 1). PFMFind client takes user input, and communicates with FSIndex and the database through its components. It passes search parameters in batches to FSIndex and receives the results of searches that are then stored in the database. It also retrieves the results from the database and displays them, together with available annotations, to the user. The annotations are stored in a separate (BioSQL) schema in the database.

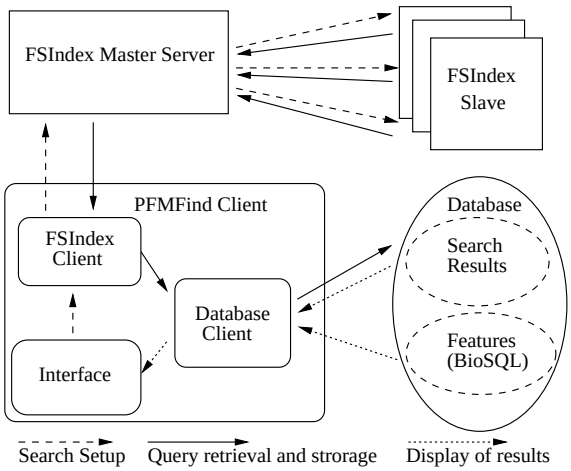

Figure 1: Structure of PFMFind system.

Most of PFMFind was written in the Python programming language, and uses both the standard Python library and additional modules such as Biopython (http://www.biopython.org). The components communicate using the standard TCP/IP socket interface and can therefore be located on different machines. Since PFMFind is highly modular, the GUI client can be replaced by a Python script for non-interactive use.

Similarity Search

PFMFind supports searches of datasets of short peptide fragments of fixed length using an ungapped similarity score obtained by summing similarity scores at each position of the fragments being compared. The positional similarity scores can be defined by standard score matrices such as PAM (3) or BLOSUM (6), or by PSSMs (position specific score matrices) (4). A dataset consists of all fragments of a specified length from a given protein sequence dataset (where the fragments may overlap).

Iterative construction of PSSM, similar to that used by PSI-BLAST (1), is supported through plugins — Python routines that take the results of a previous search and construct a PSSM. The default plugin uses the weighting procedure of Henikoff and Henikoff (5) to assign weights to

fragments and Dirichlet mixtures (9) for regularising the amino acid frequency counts at each position. Users with some knowledge of Python can create their own plugins and use them for searches by placing then in the appropriate directory.

Search criteria can be specified according to cutoff raw similarity scores, distances, p-values and E-values, as well as the number of closest datapoints to retrieve. The probability model for calculation of p-values assumes that the score of each fragment is the sum of independent random variables corresponding to the score at each position and the score distribution is calculated using FFT.

FSIndex

The heart of PFMFind is FSIndex, an efficient indexing scheme for similarity search of very large datasets of short protein fragments of fixed length (8; 10). FSIndex is based on two principles: reduction of the amino acid alphabet to clusters largely based on their biochemical properties (hydrophobic, polar, charged, aromatic ...) and combinatorial generation of neighbours. The design of FSIndex means that a typical search involves scanning less than 1% of the fragment dataset, but ensures that no neighbours satisfying search criteria are ever missed.

FSIndex is implemented in the C programming language and embedded into Python, with the whole data structure as well as the indexed sequences stored in primary memory. For even greater efficiency, computation of searches can be distributed among several machines using a master/slave model: the master handles p-value computations, distributes queries to slaves, each of which is indexing a different part of the dataset, and communicates with the client.

Database

The second major component of PFMFind is a relational database, used both for storage of search results and the sequence annotation. We use PostgreSQL, a freely available modern database management system.

Each user of the system has their own schema for storing search results. The database also stores all search parameters, including PSSMs and the results of each iteration, facilitating reversion to a previous iteration without repeating the whole procedure.

The database stores sequence annotations in a standard BioSQL schema available to all users. PFMFind also contains scripts for loading four types of information beyond the basic sequence information: Uniprot (2) keywords and features, Uniref clusters (2) and InterPro (7) domains. When retrieved for display, annotations are joined to search results through accession numbers.

GUI Client

The final PFMFind component is a GUI client that connects to both the FSIndex master and the database component. To perform fragment searches, the user specifies a query sequence, usually a long sequence that is broken into overlapping fragments of fixed length, and chooses the fragment lengths, cutoff parameters and the actual fragments in the query sequences that will be used for the search.

The GUI client can display search results both as lists of hits associated with a particular location in the query sequence and as a feature vs location dot plot — each location matching a particular feature is marked by a coloured dot. Dots are colour coded by the number of hits matching the feature to distinguish frequently represented features from those that appear only a few times in the hit list. All computations of PSSMs are performed by the GUI client as well.

Conclusion

PFMFind is an efficient, flexible, and extensible framework for similarity search of datasets of short peptide fragments. It supports fast similarity search with selectivity and sensitivity specified by PSSMs and associates search results with biological function by using sequence features and annotations. We shall describe our use of PFMFind to search for functions associated with short fragments that could have arisen by convergence in another publication.

Acknowledgements

We wish to thank Pavle Mogin and Danyl McLauchlan for their help with PostgreSQL and testing the software, respectively. A.S. was supported by a Bright Future PhD scholarship awarded by the NZ Tertiary Education Commission jointly with the Fonterra Research Centre and by a Fields Institute/University of Ottawa postdoctoral fellowship. V.G.P. and A.S. acknowledge support from NSERC discovery grant RGPIN/261450-2003 and University of Ottawa internal grants.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[2] Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 33 Database Issue:154–159, 2005.

[3] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, chapter 22, pages 345–352. National Biomedical Research Foundation, 1978.

[4] M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 84:4355–4358, 1987.

[5] S Henikoff and J G Henikoff. Position-based sequence weights. *J. Mol. Biol.*, 243(4):574–578, 1994.

[6] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919, 1992.

[7] Nicola J Mulder, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Paul Bradley, Peer Bork, Phillip Bucher, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Richard Durbin, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicola Harte, Nicolas Hulo, Daniel Kahn, Alexander Kanapin, Maria Krestyaninova, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Jennifer McDowall, Alex Mitchell, Anastasia N Nikolskaya, Sandra Orchard, Marco Pagni, Chris P Ponting, Emmanuel Quevillon, Jeremy Selengut, Christian J A Sigrist, Ville Silventoinen, David J Studholme, Robert Vaughan, and Cathy H Wu. InterPro, progress and status in 2005. *Nucleic Acids Res.*, 33 Database Issue:201–205, 2005.

[8] Vladimir Pestov and Aleksandar Stojmirović. Indexing Schemes for Similarity Search: An Illustrated Paradigm. *Fund. Inform.*, 70:367–385, 2006.

[9] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, 12(4):327–345, 1996.

[10] Aleksandar Stojmirović and Vladimir Pestov. Indexing schemes for similarity search in datasets of short protein fragments. ArXiv e-print cs.DS/0309005, version 2, Jan 2006, 14 pp.

[11] Terry J Watt and Donald F Doyle. ESPSearch: a program for finding exact sequences and patterns in DNA, RNA, or protein. *Biotechniques*, 38(1):109–115, 2005.